

# Learn-As-You-Go Acceleration of Cosmological Parameter Estimates

Grigor Aslanyan, Richard Easther, and Layne C. Price

Department of Physics, University of Auckland, Private Bag 92019, Auckland, New Zealand

E-mail: [g.aslanyan@auckland.ac.nz](mailto:g.aslanyan@auckland.ac.nz), [r.easther@auckland.ac.nz](mailto:r.easther@auckland.ac.nz),  
[lpri691@aucklanduni.ac.nz](mailto:lpri691@aucklanduni.ac.nz)

**Abstract.** Cosmological analyses can be accelerated by approximating slow calculations using a training set, which is either precomputed or generated dynamically. However, this approach is only safe if the approximations are well understood and controlled. This paper surveys issues associated with the use of machine-learning based emulation strategies for accelerating cosmological parameter estimation. We describe a learn-as-you-go algorithm that is implemented in the COSMO++ code<sup>1</sup> and (1) trains the emulator while simultaneously estimating posterior probabilities; (2) identifies unreliable estimates, computing the exact numerical likelihoods if necessary; and (3) progressively learns and updates the error model as the calculation progresses. We explicitly describe and model the emulation error and show how this can be propagated into the posterior probabilities. We apply these techniques to the *Planck* likelihood and the calculation of  $\Lambda$ CDM posterior probabilities. The computation is significantly accelerated without a pre-defined training set and uncertainties in the posterior probabilities are subdominant to statistical fluctuations. We have obtained a speedup factor of 6.5 for Metropolis-Hastings and 3.5 for nested sampling. Finally, we discuss the general requirements for a credible error model and show how to update them on-the-fly.

---

<sup>1</sup><http://cosmopp.com>

---

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Emulation Algorithm</b>	<b>3</b>
<b>3</b>	<b>Error Analysis</b>	<b>6</b>
<b>4</b>	<b>Learn-As-You-Go Techniques</b>	<b>8</b>
<b>5</b>	<b>Propagating Local Errors to Posteriors</b>	<b>9</b>
<b>6</b>	<b>CMB and <i>Planck</i> Likelihood</b>	<b>12</b>
6.1	Details of procedure	12
6.2	Emulated $\Lambda$ CDM posteriors	13
6.3	Estimated improvement	15
<b>7</b>	<b>Conclusion</b>	<b>18</b>
<b>A</b>	<b>Least Squares Fit</b>	<b>19</b>
<b>B</b>	<b>Implementation in Cosmo++</b>	<b>20</b>
B.1	$k$ -d tree	20
B.2	Emulator	21
B.3	Error Model	22
B.4	Learn-As-You-Go	23
B.5	Accelerated <i>Planck</i> Likelihood	23

---

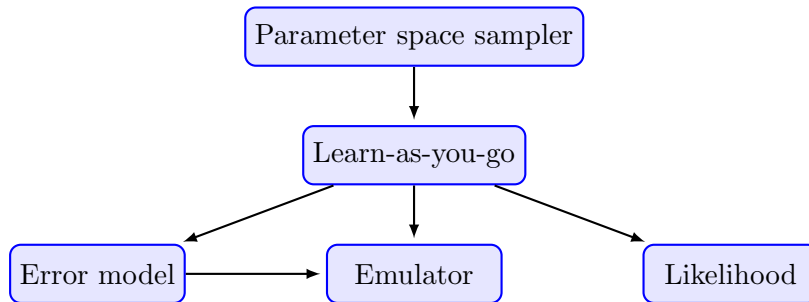
## 1 Introduction

Constraining cosmological models using experimental data typically requires many evaluations of the data likelihood, which can be very slow. For instance, the analysis of galaxy surveys typically requires an estimate of the non-linear matter power spectrum and using  $N$ -body simulations to achieve the desired accuracy would be computationally prohibitive. Even calculating data likelihoods for cosmic microwave background (CMB) data, such as WMAP [1] and *Planck* [2], requires the angular power spectra, which must be obtained through the evolution of Boltzmann codes like CAMB [3] and CLASS [4, 5] that take significant fractions of a second to calculate for each set of parameters in a long Markov chain.

A common approach to speeding up computationally expensive calculations is to employ a statistical model that *emulates* an expensive calculation using a *training set* from which results are empirically approximated.<sup>1</sup> Several such algorithms for cosmological data analysis have been described, and associated codes released [7–10]. Recombination and the CMB likelihood calculations are emulated by RICO [11] and PICO [12, 13] respectively, while the likelihoods themselves are cached and re-used for interpolation with INTERPMC [14]. The

---

<sup>1</sup>Following the statistics literature, *e.g.*, Ref. [6], we use the term *emulator* to refer to a statistical or deterministic model that approximates the output of a *simulation* code in a computationally efficient fashion.



**Figure 1.** The relationship between different classes in the object-oriented implementation of our learn-as-you-go algorithm. An arrow pointing from one block to another indicates the first class “knows about” or uses the output of the second.

COYOTE UNIVERSE suite [15–21] emulates the matter power spectrum using Gaussian processes and  $N$ -body simulations, with significant attention placed on the optimal construction of training sets.

Although emulators can significantly accelerate calculations, it can be difficult to estimate and control the errors they induce. Consequently, emulators can only be used if the methodology is known to be highly reliable. Two obvious but unattractive solutions to this dilemma are to make direct comparisons with the final exact calculation or to build very dense training sets. Further, because the specific regions of parameter space to be explored may not be known in advance, the training must either span a large region of parameter space or be extended dynamically in order to ensure accuracy. Obviously, the advantages of emulation are eroded if significant computational resources are needed to train the emulator or the uncertainty associated with the emulation is not well-controlled.

An approach to solving this problem was demonstrated in Refs [14, 22–24], where an external statistical sampling method was coupled to a machine learning algorithm to dynamically refine a statistical emulator. This results in a training set that is dense only in areas of parameter space where it is needed, while simultaneously evaluating a model’s posterior. We call methods of this sort *learn-as-you-go emulators*, since they both accelerate a calculation that would be performed in any case and build a training set for future use.

In this paper we introduce a flexible and powerful learn-as-you-go algorithm that incorporates an adaptive and trainable error model for the differences between emulated and “exact” cosmological likelihood evaluations.<sup>2</sup> We use well-tested sampling methods to simultaneously evaluate model posteriors, generate the training set, and update the error model by cross-validation. Critically, the local emulation errors are propagated through the calculation, allowing us to estimate the error in the posterior probability  $p(\mathbf{x} | D)$  for data  $D$  and parameter  $\mathbf{x}$  induced by our use of emulation. Consequently, the uncertainty in the posteriors is modelled explicitly, quantifying the robustness of the estimated posterior.

Cosmological parameter estimation using our algorithm has five major components:

1. **Sampler.**— An externally defined method to explore parameter space.
2. **Data likelihood.**— The pre-defined function that we intend to emulate.

---

<sup>2</sup>A fully documented implementation of the algorithm is included in COSMO++ [25], available at <http://cosmopp.com>. The code is modular and could easily be applied to similar problems in other fields. In addition, we provide a fast version of the *Planck* likelihood module that uses this algorithm (see Sect. 6 and Appendix B for further details).

3. **Emulator.**— Given a training set, perform a rapid emulation for new parameter values of interest.
4. **Error model.**— For every new point, estimate the error made by emulating. Propagate the errors in the likelihood from the emulation process into errors for the posterior distributions.
5. **Learn-as-you-go.**— Continuously update the training set of the emulator based on new exact evaluations. Also, continuously update the error model. Determine whether the emulated value is acceptable or if the exact calculation should be performed.

Fig. 1 describes the relationship between the different classes in our publicly available, object-oriented C++ implementation for parameter space sampling and likelihood evaluations.

Our techniques are not restricted to cosmology and can be easily applied to any computationally expensive calculation that depends on a fixed number of input parameters. The algorithm does not make strong assumptions about the properties of the parameter space and is usable with a wide variety of statistical samplers. Frequently repeated operations take at most  $O(\log N_{\text{tr}})$  time, where  $N_{\text{tr}}$  is the size of the training set, so that the speed of the algorithm is only weakly dependent on  $N_{\text{tr}}$ . Furthermore, the cost of evaluating the error model and adding new points to the training set is negligibly small for typical calculations. Our implementation scales well in an MPI environment, with all processes sharing a single training set. Unlike the approach of Ref. [14] we do not assume that the parameter space is simply connected or has a single, global maximum.

Following Ref. [14], we use the emulated value of the data likelihood for any set of parameters for which the emulation error is below a user-defined threshold; otherwise the exact calculation is performed and its result added to the training set. Defining an unacceptable level of error between emulated and non-emulated values lets the user tradeoff between accuracy and acceleration, assuming that the error model is sufficiently robust. This approach ensures that there are no redundancies in the training set: new points are not added where they are not needed and no time is spent generating training data that are not used. In common with the techniques described in Refs [14, 23], we do not need an initial training set, but providing one can significantly increase the resulting acceleration.

We demonstrate the use of the algorithm by accelerating estimates of posterior probabilities for  $\Lambda$ CDM with CMB likelihoods. CMB analyses provide a useful testing platform for our methods because they can be conducted in a reasonable time but are still slow enough that any improvement in speed will be extremely welcome. Starting with an empty training set we use the emulator with two parameter space samplers: the Metropolis-Hastings [26] MCMC method implemented in COSMO++ [25] and nested sampling with the MULTINEST package [27–29]. We track the number of exact likelihood calculations required with and without emulation, as well as the overall speedup. We demonstrate that the use of the emulator results in negligibly small errors in the posterior distributions of the parameters as determined by both our pre-defined error model and the exact posteriors obtained without emulation.

## 2 Emulation Algorithm

While cosmological applications will be primarily concerned with emulating the likelihood function  $\mathcal{L}(D|\mathbf{x})$ , for observational data  $D$  and model parameters  $\mathbf{x}$ , in general we could

Symbol	Description	Symbol	Description
$p$	Probability distribution function	$D$	Data
$\mathcal{L}$	Data likelihood	$f$	Exact function intended for emulation
$f_{\text{em}}$	Emulating function that approximates $f$	$n$	Number of dimensions in parameter space (domain of $f$ )
$\mathbf{x}$	Parameter space vector	$X_{\text{train}}$	Training set of parameter space points
$Y_{\text{train}}$	Range of training set $f(X_{\text{train}})$	$m$	Number of dimensions for the range of $f$
$\rho$	Model-dependent measure of local sparsity of training points	$k$	Number of nearest neighbors used in analysis
$w$	Weighting of nearest neighbors	$C$	Covariance matrix of training set
$L$	Cholesky decomposed covariance $C \equiv LL^T$	$\delta$	The degree of the polynomial interpolation for $f_{\text{em}}$
$\Delta f$	Local emulation error, <i>i.e.</i> , difference between $f_{\text{em}}(\mathbf{x})$ and $f(\mathbf{x})$	$e_\eta$	Local emulation error on arbitrary scalar function $\eta$
$\Delta p$	Global emulation error in posterior probability	$\boldsymbol{\theta}_N$	Nuisance parameters
$r$	Ratio of $e_\eta$ to $\rho$		

**Table 1.** A legend for our notation.

emulate any arbitrary function  $f(\mathbf{x})$ . In this Section we will keep our discussion as general as possible, and will introduce an explicit local error model in Sect. 3.

Suppose we need to calculate the function

$$f : \mathbb{R}^n \rightarrow \mathbb{R}^m \quad (2.1)$$

a large number of times with a specified accuracy. The exact numerical evaluation of  $f$  is assumed to be slow, so we wish to store previously calculated results in order to rapidly estimate the value of  $f$  for new input points  $\mathbf{x} \in \mathbb{R}^n$ . The emulated value of  $f$  is denoted by  $f_{\text{em}}$ .

A detailed schematic of the emulation algorithm is given in Alg. 1 and the implementation of our publicly available code is discussed in Appendix B. In particular, the  $k$ -d tree used to locate the points used for the emulation is described in B.1 and the emulator implementation is discussed in B.2. We expand on each of these functions below:

**InitializeEmul.**— This function builds and saves a  $k$ -d tree from the training set  $X_{\text{train}}$  following Ref. [30]. We linearly transform the input points  $\mathbf{x} \rightarrow \mathbf{x}'$  for a new basis in which all of the parameters in the transformed training set are uncorrelated and where all of the components for each  $\mathbf{x}'$  have the same order of magnitude. To achieve this, we calculate the

---

**Algorithm 1** Emulates a function  $f$  that maps  $\mathbf{x} \in \mathbb{R}^n \rightarrow \mathbf{y} \in \mathbb{R}^m$ , given a training set consisting of parameter space points  $\{\mathbf{x}_i\}$  and their mappings  $\{\mathbf{y}_i = f(\mathbf{x}_i)\}$ .

---

**function** INITIALIZEEMUL( $X_{\text{train}} = \{\mathbf{x}_i : i = \overline{1, N_{\text{tr}}}\}, Y_{\text{train}} = \{\mathbf{y}_i : i = \overline{1, N_{\text{tr}}}\}$ )

Calculate the sample covariance matrix  $C$  of points  $\mathbf{x}_i$  in  $X_{\text{train}}$

Cholesky decompose  $C$  into matrices  $L$  by  $C \equiv LL^T$

Build training set in new basis  $X'_{\text{train}} = \{L^{-1}\mathbf{x}_i : i = \overline{1, N_{\text{tr}}}\}$

Build  $k$ -d tree  $T_{kd}$  from  $X'_{\text{train}}$

**function** EMULATE( $\mathbf{x}_0$ )

Change basis by defining  $\mathbf{x}'_0 \equiv L^{-1}\mathbf{x}_0$

Using  $T_{kd}$ , find  $k$  nearest neighbors  $X^{\text{NN}} = \{\mathbf{x}_i^{\text{NN}} : i = \overline{1, k}\}$  of  $\mathbf{x}'_0$

Find the corresponding  $Y^{\text{NN}} \equiv \{\mathbf{y}_i^{\text{NN}} = f(\mathbf{x}_i^{\text{NN}}) : i = \overline{1, k}\}$  from  $Y_{\text{train}}$

**for all** nearest neighbors  $\mathbf{x}_i^{\text{NN}} \in X^{\text{NN}}$  **do**

Calculate weighting  $w^i \equiv 1/\text{Distance}(\mathbf{x}'_0, \mathbf{x}_i^{\text{NN}})$  for some metric

**for all**  $j = \overline{1, m}$  **do**

Assume  $(f_{\text{em}}(\mathbf{x}))_j \equiv P_j(\mathbf{x}')$  around  $\mathbf{x}'_0$  where  $P_j$  is a polynomial

Find coefficients of  $P_j$  to minimize  $S_j = \sum_{i=1}^k w^i |P_j(\mathbf{x}_i^{\text{NN}}) - (\mathbf{y}_i^{\text{NN}})_j|^2$

**return**  $f_{\text{em}}(\mathbf{x}) = (P_1(\mathbf{x}'_0), \dots, P_m(\mathbf{x}'_0))$

**function** ADDPOINTTOTRAINSET( $\mathbf{x}, \mathbf{y}$ )

Add  $\mathbf{x}$  to  $X_{\text{train}}$  and  $\mathbf{y}$  to  $Y_{\text{train}}$

Add  $\mathbf{x}' \equiv L^{-1}\mathbf{x}$  to  $X'_{\text{train}}$  and  $T_{kd}$

**if**  $\text{Depth}(T_{kd}) \geq 4 \log N_{\text{tr}}$  **then**

Rebalance  $T_{kd}$

**procedure** SIMPLEEMULATOR

Calculate training sets  $X_{\text{train}}$  and  $Y_{\text{train}} = f(X_{\text{train}})$  exactly

Set up the emulator by calling INITIALIZEEMUL( $X_{\text{train}}, Y_{\text{train}}$ )

Emulate  $f(\mathbf{x})$  by calling EMULATE( $\mathbf{x}$ )

---

sample correlation matrix,  $C$ , of the points in the training set, find its Cholesky decomposition matrix  $C = LL^T$ , and use  $L$  to change the basis to  $\mathbf{x}' \equiv L^{-1}\mathbf{x}$ .

**Emulate.**— This function computes  $f_{\text{em}}$  for an input point  $\mathbf{x}_0$ , finding the  $k$  nearest neighbors ( $k$ -NN) of  $\mathbf{x}$  in the  $k$ -d tree built from  $X_{\text{train}}$ . These near-by points in the training set are then interpolated and the value of the polynomial interpolation at  $\mathbf{x}_0$  is used for  $f_{\text{em}}(\mathbf{x}_0)$ .

As the  $k$ -d tree in INITIALIZEEMUL stores input points in the Cholesky basis,  $\mathbf{x}_0$  is first transformed to that basis with  $\mathbf{x}_0 \rightarrow \mathbf{x}'_0$ , after which the  $k$  nearest neighbors are located using the  $k$ -d tree. Given we are working with dimensionless, normalized variables we use the Euclidean distance to put a metric on the parameter space, but other possibilities exist including the Fisher information metric. In general,  $f(\mathbf{x}) = (f_1, \dots, f_m)$  with  $m \geq 1$ , so for each dimension  $j = 1, \dots, m$  we build the  $j^{\text{th}}$  component of  $f_{\text{em}}$  by an independent,  $n$ -dimensional polynomial interpolation near the point  $\mathbf{x}'_0$ . The coefficients of the polynomial are found simply by doing a weighted least squares fit to the  $k$  nearest neighbors, with weights  $w^i$  inversely proportional to the Euclidean distance. The details of the fit are discussed in Appendix A.

An  $n$ -dimensional polynomial of degree  $\delta$  will have  $N_{\text{coeff}}$  coefficients, with

$$N_{\text{coeff}} = 1 + n + \frac{n(n+1)}{2} + \dots + \frac{n(n+1) \cdots (n+\delta-1)}{\delta!}. \quad (2.2)$$

To avoid degeneracies we use  $k = 2N_{\text{coeff}}$ . It is easy to see that  $N_{\text{coeff}}$  increases exponentially with  $\delta$ , so choosing a high dimensional polynomial will require a large number of nearest neighbors, as well as a least squares fit in a high-dimensional space, resulting in a uselessly slow emulator. Consequently, we have restricted our implementation to linear and quadratic polynomials.

**AddPointToTrainSet.**— This function adds new points to the training set by first transforming to the Cholesky basis and then adding them to the appropriate location in the  $k$ -d tree. After adding many points the  $k$ -d tree may become unbalanced, resulting in a slowdown of the nearest neighbor search. After adding a new point we check if the tree has become unbalanced, and rebuild it if necessary. The criterion for this step is that the depth of the tree is four times larger than the depth of a perfectly balanced tree. This ensures that the tree always has an  $O(\log N_{\text{tr}})$  depth. In practice, this rebuilding is sufficiently infrequent that it makes a negligible contribution to the overall computational cost of the algorithm.

**SimpleEmulator.**— This is the simplest possible implementation of the above functions. After obtaining a training set, INITIALIZEEMUL builds the  $k$ -d tree with  $f_{\text{em}}(\mathbf{x})$  computed by EMULATE for a parameter space point  $\mathbf{x}$ .

**Computational complexity.**— If the dimensions  $n$  and  $m$  of the input and output spaces of  $f_{\text{em}}$ , respectively, are much smaller than  $N_{\text{tr}}$ , the run-time of the algorithm depends primarily on the size  $N_{\text{tr}}$  of the training set. The complexity of INITIALIZEEMUL is dominated by the construction of the  $k$ -d tree, which is  $O(N_{\text{tr}} \log N_{\text{tr}})$ . Finding the  $k$  nearest neighbors takes only  $O(\log N_{\text{tr}})$  time [30] and the interpolation step is independent of  $N_{\text{tr}}$ , which fixes the complexity of EMULATE. Finally, adding a new point to the  $k$ -d tree is linear in the depth of the tree, *i.e.*, adding new points is  $O(\log N_{\text{tr}})$ . However, the infrequent rebalancing step scales as  $O(N_{\text{tr}} \log N_{\text{tr}})$ .

### 3 Error Analysis

We emulate the function  $f$  in Eq. (2.1) with an approximate function  $f_{\text{em}}$  using the method described in Sect. 2. However, we also want to estimate the difference between  $f$  and  $f_{\text{em}}$ . This is critical both to the estimate of the total error in our calculation and when determining whether or not the approximation  $f_{\text{em}}(\mathbf{x})$  is acceptable at a specific point  $\mathbf{x}$ .

We define the local emulation error  $\Delta f$  at a point  $\mathbf{x}$  as

$$\Delta f(\mathbf{x}) \equiv f_{\text{em}}(\mathbf{x}) - f(\mathbf{x}). \quad (3.1)$$

Unless  $\mathbf{x}$  is an element of the training set  $X_{\text{train}}$  we will not know  $f(\mathbf{x})$ , so we treat  $\Delta f$  as a random variable with a probability distribution  $p(\Delta f)$ . It is also convenient to define a local error  $e_\eta$  via a scalar function  $\eta : \mathbb{R}^m \rightarrow \mathbb{R}$ ,

$$e_\eta(\mathbf{x}) = \eta(f_{\text{em}}(\mathbf{x})) - \eta(f(\mathbf{x})). \quad (3.2)$$

---

**Algorithm 2** Building an empirical model for the local error  $e_\eta$  on the scalar function  $\eta$ , using a set of exact calculations  $Y_{CV} = f(X_{CV})$  that the emulating function  $f_{em}$  has not yet been trained on.

---

**function** BUILDERRORMODEL( $\eta, X_{CV} = \{\mathbf{x}_i : i = \overline{1, M}\}, Y_{CV} = \{\mathbf{y}_i = f(\mathbf{x}_i) : i = \overline{1, M}\}$ )  
 Initialize an empty list  $R$   
**for all** points  $\mathbf{x}_i$  in  $X_{CV}$  **do**  
   Calculate  $\rho(\mathbf{x}_i)$ , *e.g.*, as the mean distance to the  $k$  nearest neighbors  
   Add ratio  $|\eta(\text{EMULATE}(\mathbf{x}_i)) - \eta(\mathbf{y}_i)|/\rho(\mathbf{x}_i)$  to  $R$ , with EMULATE from Alg. 1  
 Create histogram  $H$  from  $R$   
 Smooth and normalize  $H$  to obtain an estimate of  $p(|e_\eta|/\rho)$

**procedure** SIMPLEERROR

  Calculate sets  $X_{CV}$  and  $Y_{CV} = f(X_{CV})$  exactly  
   Set up the emulator by calling INITIALIZEEMUL( $X_{\text{train}}, Y_{\text{train}}$ )  
   Estimate  $p(|e_\eta|/\rho)$  by calling BUILDERRORMODEL( $\eta, X_{CV}, Y_{CV}$ )  
   For point  $\mathbf{x}$  calculate  $p(e_\eta(\mathbf{x})) \propto \rho(\mathbf{x}) \times p(|e_\eta|/\rho)$  and renormalize

---

The form of  $\eta$  is not unique, beyond requiring that it reduces the dimensionality of the output space of  $f$  to unity. Possible choices include a spatial average on the output space,  $\eta(f(\mathbf{x})) \equiv \langle f(\mathbf{x}) \rangle$ , or by evaluating the likelihood for experimental data,  $\eta(f(\mathbf{x})) \equiv p(D | f(\mathbf{x}))$ .<sup>3</sup>

We want an error model that is broadly applicable, computationally efficient, and updates automatically as the training set grows. We expect that  $p(\Delta f)$  and  $p(e_\eta)$  will be strongly dependent on the distances between  $\mathbf{x}$  and its nearest points in  $X_{\text{train}}$ , which are used for the interpolation. We will use a single parameter  $\rho(\mathbf{x} | X_{\text{train}})$  to define the probability distribution  $p(e_\eta | \rho, X_{\text{train}})$ . We choose  $\rho$  to be inversely proportional to the density of the training set  $X_{\text{train}}$  near  $\mathbf{x}$ , typically the mean  $n$ -dimensional Euclidean distance to the  $k$  nearest neighbors in  $X_{\text{train}}$ .

The form of  $p(e_\eta | \rho, X_{\text{train}})$  could be modelled using *a priori* knowledge of  $f$  and  $\eta$ , but it is far more flexible to estimate  $p(e_\eta | \rho, X_{\text{train}})$  empirically, via cross-validation on subsets  $X_{CV}$  of  $X_{\text{train}}$ , from which  $e_\eta$  can be calculated exactly. We assume that  $e_\eta$  increases linearly with  $\rho$ , allowing us to define a probability distribution on their ratio as

$$p(e_\eta | \rho, X_{\text{train}}) \rightarrow p\left(\frac{e_\eta}{\rho} | X_{CV}\right), \quad (3.3)$$

up to an arbitrary normalization.

We describe our error model algorithm schematically in Alg. 2 and its detailed implementation is given in Appendix B.3. The function BUILDERRORMODEL takes as input the cross validation sets  $X_{CV}$  and  $Y_{CV} = f(X_{CV})$ . For each point  $\mathbf{x}_i$  in  $X_{CV}$  it calculates the local scalar error  $e_\eta(\mathbf{x}_i)$  and the training set sparsity measure  $\rho(\mathbf{x}_i)$ , using the mean Euclidean distance to the  $k$  nearest neighbors. We enforce our assumption  $e_\eta \propto \rho$  by constructing a smoothed and normalized histogram for the ratio  $|e_\eta|/\rho$ , from which we empirically estimate  $p(|e_\eta|/\rho)$ .<sup>4</sup> SIMPLEERROR evaluates  $p(e_\eta(\mathbf{x}))$  by inverting the transformation in Eq. (3.3),

---

<sup>3</sup>We will typically assume that  $\eta(\mathbf{y})$  is inexpensive to calculate compared to  $f(\mathbf{x})$ , if we are given  $\mathbf{y}$ . If it is not, then we would ideally emulate the composite function  $f' = \eta \circ f$  directly.

<sup>4</sup>We also save the  $1\sigma$  upper bound  $b_{1\sigma}$  on  $p(|e_\eta|/\rho)$  for use in the learn-as-you-go algorithm.



---

**Algorithm 3** Implementation of emulation scheme (Alg. 1) and error model (Alg. 2) into a learn-as-you-go method that progressively builds a training set  $X_{\text{train}}$ , learns the error distribution  $p(e_\eta)$ , and decides when to emulate or exactly evaluate  $f(\mathbf{x})$ .

---

```

function PREPAREEMULATOR( $\eta, X_{\text{train}}, Y_{\text{train}}, \phi_{\text{CV}}$ )
  if Size( $X_{\text{train}}$ )  $\geq N_{\text{min}}$  then
    Move a fraction  $\phi_{\text{CV}}$  of the points in  $X_{\text{train}}, Y_{\text{train}}$  to  $X_{\text{CV}}, Y_{\text{CV}}$  (chosen randomly)
    Initialize the Emulator with INITIALIZEEMUL( $X_{\text{train}}, Y_{\text{train}}$ )
    Build the Error Model with BUILDErrorMODEL( $\eta, X_{\text{CV}}, Y_{\text{CV}}$ )
    Add  $X_{\text{CV}}$  and  $Y_{\text{CV}}$  back to  $X_{\text{train}}$  and  $Y_{\text{train}}$ , respectively
    Save Size( $X_{\text{train}}$ ) as CurrentSize

function EMULATIONISVALID( $\mathbf{x}$ )
  Calculate  $\rho(\mathbf{x})$ , e.g., as the mean distance to the  $k$  nearest neighbors
  Define  $e_\eta^{\text{max}} \equiv b_{1\sigma}\rho$  where  $b_{1\sigma}$  is the  $1\sigma$  upper bound of histogram  $H$  from error model
  return true if  $e_\eta^{\text{max}} < \epsilon$ , else return false

function CALCULATEEXACT( $\mathbf{x}$ )
  Add  $\mathbf{x}$  to  $X_{\text{train}}$  and the exact value  $\mathbf{y} = f(\mathbf{x})$  to  $Y_{\text{train}}$ 
  if Emulator and Error model are initialized then
    Add  $\mathbf{x}, \mathbf{y}$  to training set using ADDPOINTTOTRAINSET( $\mathbf{x}, \mathbf{y}$ )
    if Size( $X_{\text{train}}$ )  $\geq (1 + \phi_{\text{ERR}}) \times \text{CurrentSize}$  then
      Reset the Emulator and the Error Model to uninitialized
      Reinitialize them using PREPAREEMULATOR( $\eta, X_{\text{train}}, Y_{\text{train}}, \phi_{\text{CV}}$ )
  else
    if Size( $X_{\text{train}}$ )  $\geq N_{\text{min}}$  then
      Initialize using PREPAREEMULATOR( $\eta, X_{\text{train}}, Y_{\text{train}}, \phi_{\text{CV}}$ )
  return  $\mathbf{y}$ 

function CALCULATELAYG( $\mathbf{x}$ )
  if Emulator and Error Model are initialized then
    if EMULATIONISVALID( $\mathbf{x}$ ) then return EMULATE( $\mathbf{x}$ )
    else return  $\mathbf{y}_{\text{exact}} \equiv \text{CALCULATEEXACT}(\mathbf{x})$ 
  else
    return  $\mathbf{y}_{\text{exact}} \equiv \text{CALCULATEEXACT}(\mathbf{x})$ 

procedure LEARNASYOUGO
  Define an external statistical sampler STATSAMPLER, e.g., Metropolis-Hastings
  for  $\mathbf{x}$  drawn from STATSAMPLER do
    Evaluate and save the output from CALCULATELAYG( $\mathbf{x}$ ) as  $\mathbf{y}$ 

```

---

multiplying  $p(|e_\eta|/\rho)$  by  $\rho(\mathbf{x})$ , allowing for  $e_\eta$  to be positive or negative with equal probability, and renormalizing.

## 4 Learn-As-You-Go Techniques

We now combine the emulator from Sect. 2 and the error model from Sect. 3 into our learn-as-you-go algorithm. We assume the use of an externally defined statistical sampler, e.g., Metropolis-Hastings or nested sampling. The details of the emulation methods are completely

independent of the sampling strategy. The resulting algorithm is outlined schematically in Alg. 3; its detailed implementation is discussed in Appendix B.4; and we describe the purpose of its separate functions below.

**PrepareEmulator.**— If the size of the training set  $X_{\text{train}}$  exceeds a pre-defined threshold  $N_{\text{min}}$ , this function initialises the emulator. PREPAREEMULATOR is called multiple times: first when the training set size reaches  $N_{\text{min}}$  and then every time the training set size grows by a fraction  $\phi_{\text{ERR}}$ , as explained below.

PREPAREEMULATOR first divides the training set  $X_{\text{train}}$  into a smaller training set  $X'_{\text{train}} \subset X_{\text{train}}$  and a cross validation set  $X_{\text{CV}}$  by randomly choosing a fraction  $\phi_{\text{CV}}$  of points from  $X_{\text{train}}$  and removing them from the training set. It is important to ensure that points in  $X_{\text{CV}}$  do not coincide with points in  $X_{\text{train}}$ , since points in  $X_{\text{train}}$  have no emulation error by definition. The remainder of the training set is temporarily used to initialize the emulator, while  $X_{\text{CV}}$  is used to initialize the error evaluation module.

After computing the error model from  $X_{\text{CV}}$ , the points in this set are added back to  $X_{\text{train}}$ ; further evaluations rely on the complete training set  $X_{\text{train}}$ . Finally, CURRENTSIZE stores the size of  $X_{\text{train}}$  at the point when the emulator and the error model were last updated. When the size of  $X_{\text{train}}$  increases by the fraction  $\phi_{\text{ERR}}$  compared to CURRENTSIZE, we update the emulator and error model.

**EmulationIsValid.**— This function determines whether or not the approximate value  $f_{\text{em}}(\mathbf{x})$  should be accepted at a given point  $\mathbf{x}$ . Following Alg. 2, we determine the value  $b_{1\sigma}$  of the  $1\sigma$  upper bound on the probability distribution function for the error ratio  $p(|e_\eta|/\rho)$ , for the scalar function  $\eta$ . For a point  $\mathbf{x}$ , EMULATIONISVALID calculates the average Euclidean distance  $\rho(\mathbf{x})$  to its  $k$  nearest neighbors in  $X_{\text{train}}$ , then multiplies  $b_{1\sigma}$  with  $\rho(\mathbf{x})$  to determine the  $1\sigma$  upper bound on  $e_\eta(\mathbf{x})$ . Emulation is accepted if this upper bound on the error,  $e_\eta^{\text{max}}$ , is less than a target error threshold  $\epsilon$ .

**CalculateLAYG and CalculateExact.**— As in LEARNASYOUGO, CALCULATELAYG is called for any evaluation of  $f_{\text{em}}(\mathbf{x})$ . This function first checks if the error in the emulation scheme is acceptable. If it is, then the emulator from Alg. 1 is used to evaluate  $f_{\text{em}}(\mathbf{x})$ . If the error is not acceptable, CALCULATEEXACT is called to exactly evaluate  $f(\mathbf{x})$  and add  $\mathbf{x}$  to  $X_{\text{train}}$  and  $f(\mathbf{x})$  to  $Y_{\text{train}}$ . Every time a new point is added to  $X_{\text{train}}$ , this function checks if the size of  $X_{\text{train}}$  has grown by a fraction  $\phi_{\text{ERR}}$  since the error model was last updated, in which case PREPAREEMULATOR is called.

**LearnAsYouGo.**— Given a pre-defined method for obtaining points  $\mathbf{x}$  from some probability distribution of interest, CALCULATELAYG is called on  $\mathbf{x}$  and its results are saved in a Markov chain.

## 5 Propagating Local Errors to Posteriors

Let us now specialize to the important case where the function  $f(\mathbf{x})$  in Eq. (2.1), which we intend to emulate, is the logarithm of a likelihood function

$$f(\mathbf{x}) \rightarrow \ell(\mathbf{x}) \equiv \log \mathcal{L}(D | \mathbf{x}), \quad (5.1)$$

which is used by a statistical sampler to obtain posterior probability distributions. The learn-as-you-go Alg. 3 allows for estimation of the local error  $\Delta\ell$  at each point  $\mathbf{x}$ , and we will

translate this into a global uncertainty  $\Delta p(\mathbf{x} | D)$  in the posterior probability distributions. In this section we will explicitly calculate this error and in Sect. 6 we will describe its use in cosmological parameter estimation problems.<sup>5</sup>

The posterior probability distribution of the parameter space vector  $\mathbf{x}$  is related to the likelihood function  $\mathcal{L}(D | \mathbf{x})$  by Bayes theorem:

$$p(\mathbf{x} | D) = \frac{1}{Z(D)} \mathcal{L}(D | \mathbf{x}) p(\mathbf{x}), \quad (5.2)$$

where  $p(\mathbf{x})$  is the prior probability distribution for  $\mathbf{x}$ , and

$$Z(D) \equiv \int \mathcal{L}(D | \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad (5.3)$$

is the Bayesian evidence or marginalized likelihood.

We will be interested in posterior distributions for an individual parameter  $x_0$ , found after marginalizing over a set of nuisance parameters  $\boldsymbol{\theta}_N$ , given by

$$p(x_0 | D) = \int p(x_0, \boldsymbol{\theta}_N | D) d\boldsymbol{\theta}_N = \frac{1}{Z} \int \mathcal{L}(D | x_0, \boldsymbol{\theta}_N) p(x_0, \boldsymbol{\theta}_N) d\boldsymbol{\theta}_N, \quad (5.4)$$

where Bayes theorem (5.2) was used in the second relationship. The error in the emulated marginalized posterior

$$\Delta p(x_0 | D) \equiv p_{\text{em}}(x_0 | D) - p(x_0 | D) \quad (5.5)$$

can be expressed in terms of emulated quantities as

$$\Delta p(x_0 | D) = \frac{1}{Z_{\text{em}}} \int p(x_0, \boldsymbol{\theta}_N) \Delta \mathcal{L}(D | x_0, \boldsymbol{\theta}_N) d\boldsymbol{\theta}_N - \left[ \frac{p_{\text{em}}(x_0 | D)}{Z_{\text{em}}} \right] \Delta Z(D) \quad (5.6)$$

to first-order in the error terms  $\Delta \mathcal{L}$  and  $\Delta Z$ , where

$$\Delta Z(D) = \int p(\mathbf{x}) \Delta \mathcal{L}(D | \mathbf{x}) d\mathbf{x}. \quad (5.7)$$

The error in the Bayesian evidence  $\Delta Z$  and data likelihood  $\Delta \mathcal{L}$  are defined similarly to Eq. (5.5), where emulated quantities are evaluated using  $\ell_{\text{em}}$  from the learn-as-you-go emulation scheme of Sect. 4.

Assuming the errors  $\Delta p$  and  $\Delta \mathcal{L}$  are small, we can rearrange Eqs (5.6) and (5.7) as

$$\begin{aligned} \Delta \log p(x_0 | D) &= \int \Delta \log \mathcal{L}(D | x_0, \boldsymbol{\theta}_N) \mathcal{P}(\boldsymbol{\theta}_N | x_0) d\boldsymbol{\theta}_N \\ &\quad - \int \Delta \log \mathcal{L}(D | x, \boldsymbol{\theta}_N) p_{\text{em}}(x, \boldsymbol{\theta}_N | D) dx d\boldsymbol{\theta}_N, \end{aligned} \quad (5.8)$$

where we have defined the normalized probability distribution

$$\mathcal{P}(\boldsymbol{\theta}_N | x_0) = \frac{p_{\text{em}}(x_0, \boldsymbol{\theta}_N | D)}{p_{\text{em}}(x_0 | D)}. \quad (5.9)$$

---

<sup>5</sup>Our publicly available code includes the posterior distribution error estimation, allowing the user to easily obtain the uncertainties on marginalized posterior distributions resulting from the emulation process.

The first term in Eq. (5.8) is the contribution to the posterior error from the local region around  $x_0$  and the second term is the uncertainty in the overall normalization for  $p(x_0 | D)$ . The posterior error  $\Delta p(x_0 | D)$  is proportional to the posterior itself, making the emulation error greatest near regions of large *a posteriori* probability.

The learn-as-you-go emulation algorithm in Sect. 4 yields a sample of points  $X_{\text{samp}}$  that are distributed according to  $p_{\text{em}}(\mathbf{x} | D)$ , which we will use to evaluate Eq. (5.8). Given  $X_{\text{samp}}$  we first calculate  $p_{\text{em}}(x_0 | D)$  by histogram estimation, marginalizing over all other parameters and taking only the  $x_0$  component of each element  $\mathbf{x} \in X_{\text{samp}}$ . Specifically, by taking only the  $x_0$  dimension from the elements of  $X_{\text{samp}}$  and dividing the result into  $N_{\text{bins}}$  bins, we can approximate  $p(x_0 | D)$  by the fraction of the total sample that is in the bin to which  $x_0$  belongs, with the global normalization set by requiring  $\int p(x_0 | D) dx_0 \equiv 1$ .

Since any  $x'_0$  in the same bin as  $x_0$  will have the same emulated posterior probability, we can obtain a sample of nuisance parameters  $\Theta_{\text{samp}}$  that follow  $\mathcal{P}(\boldsymbol{\theta}_N | x_0)$  by taking the  $\boldsymbol{\theta}_N$  from those points  $\mathbf{x}' \in X_{\text{samp}}$  that have their  $x'_0$  in the same bin as  $x_0$ . We then estimate the integrals in Eq. (5.8) by

$$\Delta \log p(x_0 | D) \approx \frac{1}{N_{\text{bin}}} \sum_{i=1}^{N_{\text{bin}}} \Delta \ell(D | x_0, \boldsymbol{\theta}_N^i) - \frac{1}{N_{\text{samp}}} \sum_{j=1}^{N_{\text{samp}}} \Delta \ell(D | \mathbf{x}_j) \quad (5.10)$$

for  $\mathbf{x}_j \in X_{\text{samp}}$  and  $\boldsymbol{\theta}_N^i \in \Theta_{\text{samp}}$ .

With a probabilistic error model for  $\Delta \ell(\mathbf{x})$  we can use the Lyapunov central limit theorem [31] to conclude that the result of each of the sums in Eq. (5.10) will be normally distributed, assuming that the  $\Delta \ell$  are uncorrelated for different points and both  $N_{\text{bin}}$  and  $N_{\text{samp}}$  are sufficiently large. We expect the correlations in the  $\Delta \ell$  to be small in our learn-as-you-go emulation scheme, since the emulator is only used for points where the training set is dense. Consequently, two nearby points will typically have different interpolation functions over their nearest neighbors, resulting in an essentially independent error. For each emulated point  $\mathbf{x}$  in  $X_{\text{samp}}$ , Alg. 2 will calculate the mean  $\mu_\ell(\mathbf{x})$  and variance  $\sigma_\ell^2(\mathbf{x})$  of the probability distribution for  $\Delta \ell(\mathbf{x})$ . For each point that is exactly evaluated,  $\Delta \ell(\mathbf{x}) \equiv 0$ .

Finally, the probability distribution for the error in the log-posterior approaches

$$p(\Delta \log p(x_0 | D)) \xrightarrow{\text{CLT}} \mathcal{N}[\bar{\mu}_0, \bar{\sigma}_0], \quad (5.11)$$

where  $\mathcal{N}$  is a normal distribution with mean

$$\bar{\mu}_0 \equiv \frac{1}{N_{\text{bin}}} \sum_{i=1}^{N_{\text{bin}}} \mu_\ell(x_0, \boldsymbol{\theta}_N^i) - \frac{1}{N_{\text{samp}}} \sum_{j=1}^{N_{\text{samp}}} \mu_\ell(\mathbf{x}_j) \quad (5.12)$$

and variance

$$\bar{\sigma}_0^2 \equiv \frac{1}{N_{\text{bin}}^2} \sum_{i=1}^{N_{\text{bin}}} \sigma_\ell^2(x_0, \boldsymbol{\theta}_N^i) + \frac{1}{N_{\text{samp}}^2} \sum_{j=1}^{N_{\text{samp}}} \sigma_\ell^2(\mathbf{x}_j). \quad (5.13)$$

We have defined the error model in Sect. 3 so that  $\mu_\ell = 0$ , which makes  $\bar{\mu}_0$  vanish. This gives a simple way to evaluate the error in each bin of the marginalized posterior  $p_{\text{em}}(x_0 | D)$  by combining the variances of the  $\Delta \ell$  for each point in the sample. By defining an upper limit  $\sigma_{\text{max},\ell}^2$  on the allowed acceptable variance in the error model for  $\Delta \ell$ , Eq. (5.13) bounds the variance in the error in the log-posterior as  $\sigma_0^2 \lesssim \sigma_{\text{max},\ell}^2 / N_{\text{bin}}$ , assuming  $N_{\text{bin}} \ll N_{\text{samp}}$ .

In practice we first evaluate the posterior probability distribution  $p_{\text{em}}(x_0 | D)$  from  $X_{\text{samp}}$  by histogram estimation, as well as the local error in each bin, *i.e.*, the first term in Eq. (5.10), ignoring the issue of normalization. Once the error is evaluated in each bin, we calculate the total normalization factor by summing all the bins and adding the errors of each bin according to the central limit theorem. We then divide each bin by the normalization factor. Keeping only first-order terms, we add the error coming from normalization to the error in each bin, to estimate the total error.

## 6 CMB and *Planck* Likelihood

We now apply this algorithm to CMB power spectrum and *Planck* likelihood calculations, using the 2013 likelihood code [2].<sup>6</sup> We show posterior probabilities for both emulated and non-emulated scenarios in Figs 2–3, including the Gaussian-distributed posterior error as calculated in Eq. (5.11). We compare different values of the tolerance on the cutoff for the allowed error in the log-likelihood as well as two standard sampling techniques to show the robustness of this approach. Finally, we discuss the speed-ups achieved.

### 6.1 Details of procedure

The *Planck* likelihood function is a product of several distinct independent likelihoods, the low- $l$  (COMMANDER) and high- $l$  (CAMSPEC) temperature likelihoods, the lensing likelihood, and the WMAP polarization likelihood. COMMANDER depends on the cosmological model through the CMB  $TT$  power spectrum only, while CAMSPEC is also a function of 14 nuisance parameters. The lensing likelihood takes as an input the  $TT$  and  $\Phi\Phi$  power spectra for the model, where  $\Phi$  is the lensing potential, and the WMAP polarization likelihood takes the  $TT$ ,  $EE$ ,  $TE$ , and  $BB$  spectra of the model as its input.

The CAMSPEC likelihood can be evaluated relatively quickly and is the only part of the total *Planck* likelihood that depends on the additional nuisance parameters, so we exclude it from the emulation scheme to significantly reduce the dimensionality of the training set. Our approximation algorithm thus uses only the cosmological parameters as input parameters and provides the  $C_l^{TT}$ , and COMMANDER, lensing, and polarization likelihoods. Each set of cosmological parameters will be calculated using the learn-as-you-go approximation algorithm, which will decide whether or not to use the exact numerical solution for the outputs. CAMSPEC is called with the interpolated  $C_l^{TT}$  values and the foreground parameters, completing the likelihood calculation.

The number of cosmological parameters  $n$  depends on the cosmological model, but typically  $n \sim \mathcal{O}(10)$ .<sup>7</sup> From Eq. 2.2 we choose quadratic interpolation between the nearest neighbors, for which we need  $1 + n + n(n + 1)/2$  free parameters.

We use the simple error model described in Sect. 3. We set  $\phi_{\text{ERR}} = 0.05$ , which uses 5% of the training set as a cross-validation set. However, when the cross-validation set size becomes larger than 1,000 we keep only 1,000 points. The function  $\eta$  chosen for error evaluation is the entire *Planck* likelihood function, evaluated from the output of the emulator. Specifically, the emulator calculates  $C_l^{TT}$  values, as well as the COMMANDER, lensing, and polarization likelihoods. The function  $\eta$  passes the  $C_l^{TT}$  and the best-fit values of the foreground parameters to CAMSPEC. The result of this is combined with the other likelihoods.

<sup>6</sup>Our public code can be easily used with the new *Planck* likelihood code when it becomes available.

<sup>7</sup>For  $\Lambda$ CDM the number of parameters is 6.

Param.	MCMC exact	MCMC approx.	MULTINEST exact	MULTINEST approx.
$\Omega_b h^2$	$0.02210 \pm 0.00028$	$0.02210 \pm 0.00029$	$0.02210 \pm 0.00028$	$0.02210 \pm 0.00028$
$\Omega_c h^2$	$0.1196 \pm 0.0027$	$0.1196 \pm 0.0027$	$0.1195 \pm 0.0026$	$0.1195 \pm 0.0027$
$h$	$0.680 \pm 0.012$	$0.680 \pm 0.012$	$0.0680 \pm 0.012$	$0.0680 \pm 0.012$
$\tau$	$0.089 \pm 0.013$	$0.089 \pm 0.013$	$0.089 \pm 0.013$	$0.089 \pm 0.013$
$n_s$	$0.9613 \pm 0.0074$	$0.9614 \pm 0.0075$	$0.9613 \pm 0.0071$	$0.9615 \pm 0.0074$
$A_s$	$3.087 \pm 0.025$	$3.088 \pm 0.025$	$3.087 \pm 0.026$	$3.088 \pm 0.025$
$\log Z$	—	—	$-4944.0 \pm 0.3$	$-4944.1 \pm 0.3$

**Table 2.** Parameter constraints from MCMC and MULTINEST samplers with and without using the emulator.

We choose the weighted average distance to the nearest neighbors as the measure  $\rho$  of the local sparsity of the training points.

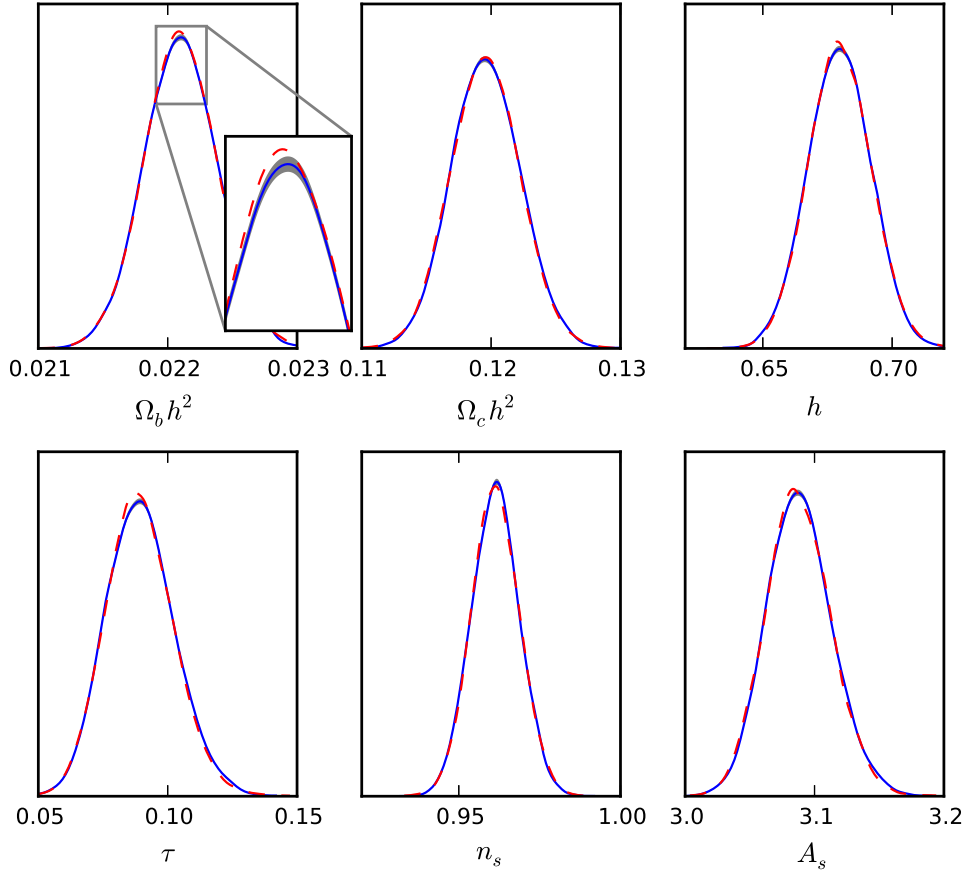
The acceptable amount of error in  $\Delta\ell$  is defined by the ratio  $r_{2\sigma}$  at which the cumulative distribution equals 0.955. For each set of input parameters we then find its  $k$  nearest neighbors in  $X_{\text{train}}$ , the weighted average Euclidean distance to them, and multiply the distance by  $r_{2\sigma}$  to find the  $2\sigma$  upper bound on the error at that point. We choose a threshold of 0.4 for the  $2\sigma$  upper bound on the error of  $-2\log\mathcal{L}$ , which corresponds to a 0.1 threshold for the  $1\sigma$  upper bound on  $\log\mathcal{L}$  for Gaussian errors. We use the emulator only if the  $2\sigma$  error at a point is below this threshold; otherwise the likelihood is calculated exactly. Emulation is not attempted until there are at least  $N_{\text{min}} = 10,000$  points in the training set.

The training set is continuously updated throughout the run, with each new point being immediately added to a  $k$ -d tree. We re-balance the tree as needed, using the criterion described in Sect. 2. We update the error model every time the training set increases in size by 25%, since likelihood evaluations for points in the cross-validation set are relatively time consuming.

We use a parallel implementation in which all threads share their training sets each time a single thread has accumulated 10 new points. When using many independent Markov chains this gives a significant performance boost, since chains that start out in different regions of the parameter space soon explore regions other chains have visited. Furthermore, we find that we frequently sample the same regions of cosmological parameter space with different values of the foreground parameters. Since these are used in the relatively fast CAMSPEC likelihood module, we have effectively implemented a fast-slow split analogous to that provided by COSMOMC [32], but without modifying the sampler itself.

## 6.2 Emulated $\Lambda$ CDM posteriors

In Fig. 2 we plot the posterior distributions for the cosmological parameters of the standard  $\Lambda$ CDM model using the *Planck* 2013 temperature data together with WMAP polarization. We use the COSMO++ MCMC sampler and compare the posterior distributions of the cosmological parameters  $\Omega_b h^2$ ,  $\Omega_c h^2$ ,  $h$ ,  $\tau$ ,  $n_s$ , and  $A_s$  with and without emulation. We also



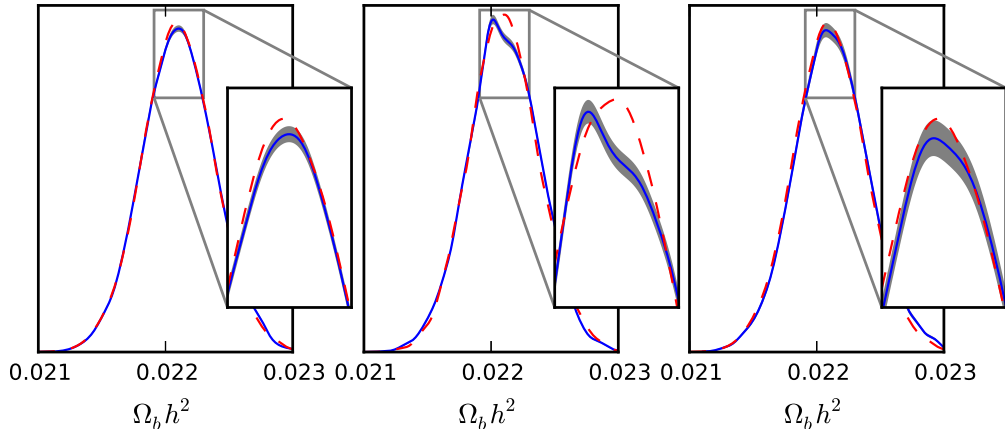
**Figure 2.** Posterior distributions of the cosmological parameters from MCMC sampling with (solid blue lines) and without (dashed red lines) emulation. The  $1\sigma$  error of the posteriors for the case of emulation is shown as a gray band around the solid blue lines.

performed this comparison for the MULTINEST sampler, with essentially identical results. Fig. 3 shows the posterior of one parameter,  $\Omega_b h^2$  obtained with MCMC (left panel) and MULTINEST (middle panel). In all cases we start with an initially empty training set.

We plot the  $1\sigma$  error range on the posteriors using the method in Sect. 5 with gray bands around the solid blue lines corresponding to the mean prediction. However, the errors are negligible and only visible if we magnify the axes substantially. Note that all posteriors include additional errors resulting from the finite sizes of the samples used to estimate the distributions, which are not included in our estimation of the errors. The uncertainty in the posterior induced by the emulation algorithm is smaller than this sampling uncertainty; and the posteriors with and without emulation agree very well.

The right panel of Fig. 3 demonstrates a more substantial error on the posterior distributions by repeating the MCMC run with only two chains and a higher threshold for the error. We have allowed points to be emulated whenever the  $2\sigma$  upper bound on  $-2\log \mathcal{L}$  is less than 1.0 instead of 0.4. Using only two chains compared to 10 decreases the number





**Figure 3.** Posterior distribution of  $\Omega_b h^2$  with (solid blue lines) and without (dashed red lines) using emulation. The  $1\sigma$  error of the posteriors for the case of emulation is shown as a gray band around the solid blue lines. The left panel is for MCMC (identical to the top left plot in Fig. 2, the middle panel is for MULTINEST, and the right panel is for MCMC again but with a larger error threshold (see the discussion in the text).

of points in the sample by a factor of 5, and increases the error bars by a factor of  $\sqrt{5}$  (see Eq. (5.13)). The error band for the posterior is wider than in the previous cases (left and middle panels) but is still comparable to the errors coming from the finite size of the sample.

The  $1\sigma$  confidence intervals for the cosmological parameters and Bayesian evidences<sup>8</sup> are summarized in Table 2. The results with emulation are almost identical to the exact case. Specifically, the confidence interval widths for the parameters vary by less than 5%, and the central values vary by less than  $\sigma/20$ . Such small numerical differences can arise even if two runs are performed with the exact same likelihood, so the emulated likelihoods involve no sacrifice of accuracy with this set of run parameters.

Figure 4 shows the estimated probability distribution of the errors  $p(|e_\eta|/\rho)$  for all the stages of error evaluation during the MCMC run. The error model is periodically updated every time the training set size increases by a fraction  $\phi_{\text{ERR}}$ , which we have chosen to be 25%. The red solid line shows the first evaluation, the blue solid line shows the last one, and the dashed lines show the intermediate stages. The increase of the training set size results in distributions with smaller expected variance for  $p(|e_\eta|/\rho)$  and a progressively higher peak at  $e_\eta/\rho = 0$ .

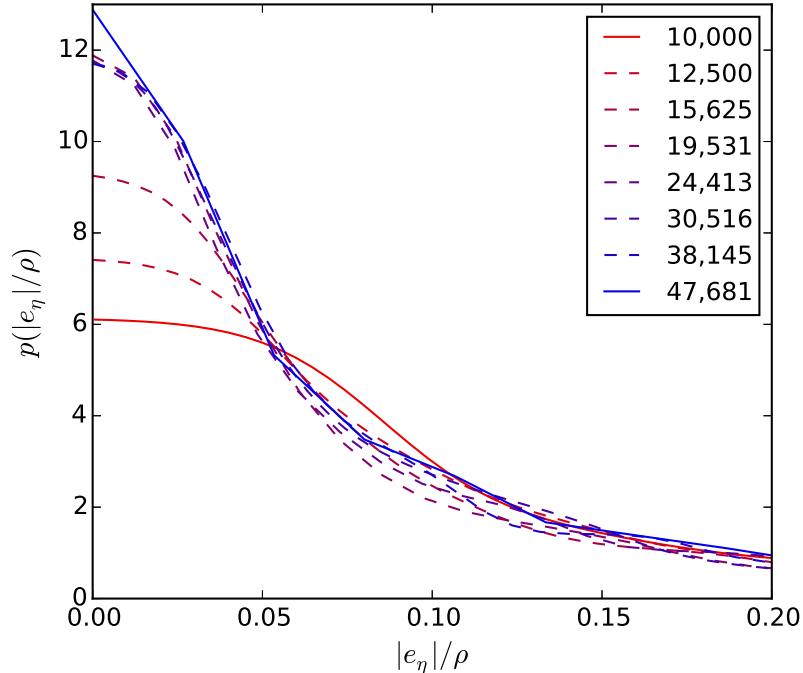
### 6.3 Estimated improvement

In Table 3 we have estimated the speedup from using the emulator for a single likelihood calculation with a well-developed training set of 150,000 points. We perform 100 exact calculations of the CMB power spectra and *Planck* likelihoods to evaluate the average run-time for each step and compare this to 100 rapid calculations using the emulator. We estimate the run-times on two different architectures, Darwin with 4 CPU cores, and Linux with 16 CPU cores, using GNU compilers on Darwin and Intel compilers with the MKL library on Linux.

---

<sup>8</sup>Computed using MULTINEST.





**Figure 4.** Probability distribution of  $|e_\eta|/\rho$  for different stages of error model evaluation (see Section 3) for the MCMC run. The labels show the size of the training set.

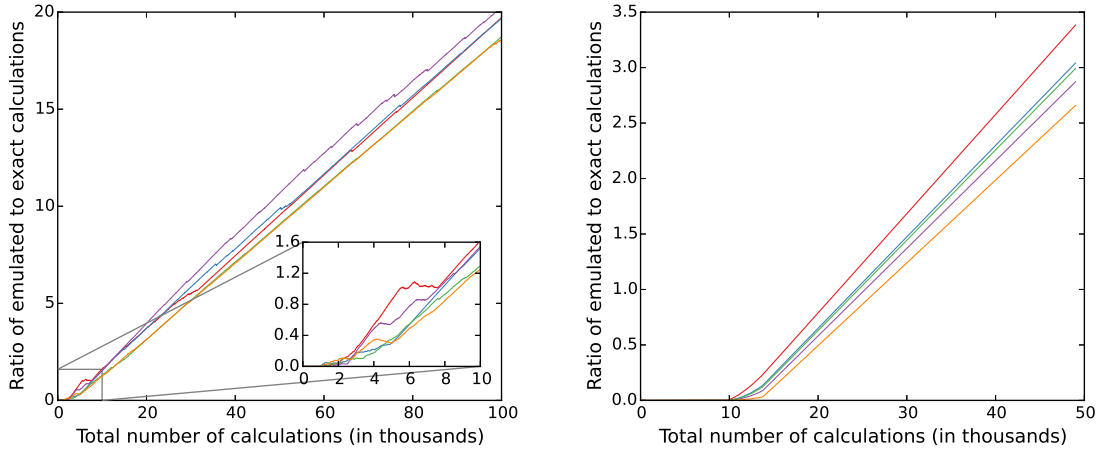
The slowest calculation is the CMB power spectrum. The polarization and lensing likelihood calculations are roughly as fast as the CMB power spectra calculation; COMMANDER and CAMSPEC are approximately one and two orders of magnitude faster, respectively. The emulated likelihood approximation is only slightly slower than CAMSPEC, meaning that the emulation itself takes a negligible amount of the run-time. Furthermore, since the computational complexity depends only logarithmically on the training set size  $N_{\text{tr}}$ , changing  $N_{\text{tr}}$  will have a negligible impact on the overall speed.

The speedups shown in Table 3 are only achievable after obtaining a large training set. However, in the learn-as-you-go scenario we start with an empty training set with the algorithm training itself with exact likelihood evaluations. Gradually the estimated error in the emulator will decrease with an increasing size for the training set and the estimated likelihood values will be reliable for new points. We will now estimate the overall speedup that can be gained using the learn-as-you-go approach separately for the MCMC and MULTINEST samplers.

We start both samplers with an empty training set and compare the results with the case of no emulation. We run MCMC using adaptive posteriors with 10 parallel chains. The burn-in length is chosen to be 500 and the Gelman-Rubin convergence diagnostic [25, 33] is used with a convergence criterion of 1.01. Convergence occurs when the chains reach a length of approximately 6,000 in both cases, with or without emulation. When no emulation is used and all of the processes are run on nodes of similar speeds the resulting chains have similar sizes. However, when using the emulator, different chains may progress at significantly different speeds depending on how frequently they walk into dense regions of the training set,

Calculation step	Linux + intel (16 cores)	Darwin + GNU (4 cores)
CMB	0.847 sec	2.703 sec
CAMSPEC	0.008 sec	0.044 sec
COMMANDER	0.062 sec	0.167 sec
Polarization	0.190 sec	1.035 sec
Lensing	0.320 sec	1.005 sec
Rapid likelihood	0.014 sec	0.050 sec
Speedup factors		
Temperature only	67	57
Temp. + pol.	81	78
Temp. + lens.	91	78
Temp. + pol. + lens.	105	99

**Table 3.** The average times of different steps of likelihood evaluation, determined from 100 repeated calculations (top). The speedup factors by using the emulator with 150,000 training points (bottom).



**Figure 5.** The ratio of the number of emulated to the number of exact likelihood evaluations as a function of the total number for the case of MCMC sampling (left) and MULTINEST sampling (right). Different colors represent 5 randomly selected chains.

resulting in chains of significantly different lengths. Convergence in this case is determined by the size of the slowest chain. MULTINEST can also be parallelized very well, so we run it with 10 MPI processes and the recommended values of the other parameters for accurate evidence calculation [28].

We plot the ratio of the number of emulated to the number of exact likelihood calcu-

lations in Fig. 5. The left panel has 5 selected MCMC chains and the right panel has 5 MPI processes for nested sampling. The different MCMC chains have significantly different lengths so we cut off all of the lines to match the smallest chain.

These plots illustrate important characteristics of our learn-as-you-go approach. At the beginning of the run the ratio of emulated points starts and remains close to zero as the emulator builds a training set. However, once the emulator achieves a high-enough accuracy the ratio of emulated points increases roughly linearly. This divides the run into “mainly training” and “mainly emulation” periods for each chain, with a transition occurring after about 5,000 total evaluations per chain for the MCMC case and after about 10,000 for MULTINEST. In the “mainly emulation” period, the ratio of emulated to exact calculations depends linearly on the total number of evaluations with a slope of  $\alpha \approx 2.11 \times 10^{-4}$  for MCMC and  $\alpha \approx 7.5 \times 10^{-5}$  for nested sampling. With a slope of  $\alpha$ , the total number of exact evaluations per chain  $N_{\text{ex}}$  will eventually saturate at  $N_{\text{ex}} = 1/\alpha$  as the total number of evaluations gets large, which is approximately  $4.8 \times 10^3$  and  $1.3 \times 10^4$  for MCMC and nested sampling, respectively. These upper limits are nearly coincident with the end of the “mainly training” period, indicating that there are almost no exactly evaluated calculations required during the “mainly emulation” stage.

MULTINEST implements nested sampling, so it starts in regions of small likelihoods and gradually converges to the peak of the likelihood [27–29]. This means that the training set is initially spread thinly across the parameter space, making emulation impossible. However, MULTINEST eventually converges on the peak of the likelihood, where the training set becomes sufficiently dense. The nested sampler then remains in the high likelihood regions, resulting in a more abrupt transition from “mainly training” to “mainly emulation” periods for MULTINEST than for MCMC, as can be seen in Fig. 5.

For MCMC sampling, the ratio of emulated to exact calculations is approximately 19 at the end of the run for the worst case chain and emulation was performed for about 95% of the points. Since the speedup factor for the temperature plus polarization likelihoods is about 80 in Table 3, an overall speedup factor of about 16 would be expected if there were no other calculations performed. In practice, the wall-clock run-times for the MCMC calculations improved by a factor of 6.5 when run with emulation and an initially empty training set. For MULTINEST, Fig. 5 shows that the final ratio of the number of emulated to the number of exact calculations reaches about 2.5–3.5 for different processes, compared to 19–20 for the MCMC sampler. For all the 10 chains, emulation was performed for 75% of all likelihood evaluations. Without any computational overhead, this would give an expected speedup of 3.9; in practice we obtained a value of 3.5.

## 7 Conclusion

We have incorporated an explicit error model in a flexible learn-as-you-go emulation algorithm. When the slow calculation  $f(\mathbf{x})$  is a data likelihood function we have also shown that the estimated local error in the log-likelihood  $\ell$  can be easily propagated into estimates of a model’s posterior probability distributions. Using a trustworthy error model then removes any need for performing redundant and slow comparisons to exact calculations.

In Sect. 2 we described an emulation algorithm that approximates an arbitrary function  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  at a parameter space point  $\mathbf{x}$  by using polynomial interpolation over the  $k$  nearest neighbors to  $\mathbf{x}$  in a pre-computed training set  $X_{\text{train}}$ . In Sect. 3 we implemented a simple error model to approximate the local difference between  $f(\mathbf{x})$  and its emulation  $f_{\text{em}}(\mathbf{x})$ .

Given an arbitrary measure  $\rho$ , which is inversely proportional to the density of the training set  $X_{\text{train}}$  at the point  $\mathbf{x}$ , our model requires the error to scale like  $\Delta f(\mathbf{x}) \sim \rho$ . We have chosen  $\rho$  to be the weighted mean Euclidean distance to the  $k$  nearest training points, although this could be relaxed in a more general error model. We empirically estimate a probability distribution  $p(\Delta f(\mathbf{x}))$  by cross-validation on a subset  $X_{\text{CV}}$  of the pre-computed training set  $X_{\text{train}}$ . Sect. 5 specializes to the case where we emulate a cosmological likelihood function  $\mathcal{L}$  and presents a method for propagating the local error  $\Delta \mathcal{L}$  to the error in marginalized posterior probability distributions. Since the error model is probabilistic, we can then report a probability distribution for the value of the posterior probability  $p(x_0 | D)$ , which tends to a normal distribution by the central limit theorem.

Improving on Ref. [14], Sect. 4 implements a pre-defined tolerance  $\epsilon$  on the allowed error  $\Delta f(\mathbf{x})$  in the emulated function. Parameter values that have emulation errors with a significant probability to be above a pre-defined threshold are directly computed, instead of emulated. Points that are calculated exactly are added to the training set  $X_{\text{train}}$  and the error model is updated repeatedly based on the increasing size and range of  $X_{\text{train}}$ . By adjusting this threshold, the algorithm emulates more or less liberally, with a corresponding trade-off between run-time improvement and accuracy. The specific details of the emulation algorithm, the error model, and the learn-as-you-go scheme can be understood from Algs 1, 2, and 3, respectively.

We have applied this methodology to  $\Lambda$ CDM parameter estimation from CMB likelihoods in Sect. 6. We chose CMB likelihoods because they are sufficiently slow to calculate that we see a moderate speed improvement when emulating, but are fast enough that we can exactly evaluate a significant fraction of the CMB Markov chains to develop a dense training set. Section 6.2 and Figs 2–3 have shown that the error in the emulation scheme is sub-dominant to the finite-sample statistical error when we choose typical cutoff values on the probability distribution of the error in  $-2 \log \mathcal{L}$ . Section 6.3 approximates a speed-up of  $\mathcal{O}(3 - 10)$  without the use of an initial training set with either MCMC or nested sampling methods. The learn-as-you-go algorithm that we have developed is parallelized, fast, and applicable to general situations outside of CMB likelihood emulation. We have publicly released all of our code in the COSMO++ package and have included the implementation details in Appendices A and B.

## Acknowledgments

The authors acknowledge the contribution of the NeSI high-performance computing facilities. New Zealand’s national facilities are provided by the New Zealand eScience Infrastructure (NeSI) and funded jointly by NeSI’s collaborator institutions and through the Ministry of Business, Innovation & Employment’s Research Infrastructure programme.<sup>9</sup>

## A Least Squares Fit

This appendix discusses the details of the weighted least squares fit of Alg. 1 in Sect. 2, which is performed by minimizing the sum

$$S_j = \sum_{i=1}^k w^i |P_j(\mathbf{x}_i^{\text{NN}}) - (\mathbf{y}_i^{\text{NN}})_j|^2. \quad (\text{A.1})$$

---

<sup>9</sup><http://www.nesi.org.nz>

The index  $j$  denotes the output coordinate of the target function  $f$  and the index  $i$  runs over all the nearest neighbors. The  $j$ -th component of  $\mathbf{y}_i^{\text{NN}}$  is  $(\mathbf{y}_i^{\text{NN}})_j$ . The weights  $w^i$  are inversely proportional to the distances of the query point  $\mathbf{x}'_0$  to the nearest neighbors  $\mathbf{x}_i^{\text{NN}}$ .

For convenience, we choose the polynomial to have the following form:

$$P_j(\mathbf{x}) = a_j + \sum_{l=1}^n b_j^l (\mathbf{x} - \mathbf{x}'_0)^l + \dots + \text{terms of degree } \delta, \quad (\text{A.2})$$

where the upper index  $l$  denotes the  $l$ -th component of a vector. Since we will only need to evaluate  $P_j$  for  $\mathbf{x}'_0$ , we will only need to calculate the value of  $a_j$ .

The coefficients of the polynomial  $P_j$  can be written as a vector

$$\mathbf{p}_j = (a_j, b_j^1, \dots, b_j^n, \dots, \text{terms of degree } \delta)^T, \quad (\text{A.3})$$

and the polynomial itself can be written in the form

$$P_j(\mathbf{x}) = \mathbf{p}_j \cdot \mathbf{z}, \quad (\text{A.4})$$

where  $\mathbf{z}$  contains all of the products of the elements of  $(\mathbf{x} - \mathbf{x}'_0)$  up to the degree  $\delta$ :

$$\mathbf{z} = (1, (\mathbf{x}_1 - (\mathbf{x}'_0)_1), \dots, (\mathbf{x}_n - (\mathbf{x}'_0)_n), \dots, \text{terms of degree } \delta). \quad (\text{A.5})$$

The values of the parameters  $\hat{\mathbf{p}}_j$  that minimize  $S_j$  can then be calculated by

$$\hat{\mathbf{p}}_j = (Z^T W Z)^{-1} Z^T W \mathbf{y}'_j, \quad (\text{A.6})$$

where

$$Z = \begin{pmatrix} \mathbf{z}_1^{\text{NN}} \\ \vdots \\ \mathbf{z}_k^{\text{NN}} \end{pmatrix}, \quad W = \begin{pmatrix} w^1 & & 0 \\ & \ddots & \\ 0 & & w^k \end{pmatrix}, \quad \mathbf{y}'_j = \begin{pmatrix} (\mathbf{y}_1^{\text{NN}})_j \\ \vdots \\ (\mathbf{y}_k^{\text{NN}})_j \end{pmatrix} \quad (\text{A.7})$$

and each  $\mathbf{z}_i^{\text{NN}}$  is calculated by substituting  $\mathbf{x}_i^{\text{NN}}$  for  $\mathbf{x}$  in (A.5). All of the matrices in Eq. (A.6) are independent of the output values of  $f$ . Consequently, the computationally costly matrix operations, including the inversion, need to be done only once even for a large dimensionality  $m$  of the output space of  $f$ .

## B Implementation in Cosmo++

The tools described in this paper have been implemented in C++ and made publicly available as a part of the COSMO++ package. In this appendix we describe the different modules available and how to use them. COSMO++ includes full documentation for all of these modules.

### B.1 $k$ -d tree

The class *KDTree* in `kd_tree.hpp` implements a simple  $k$ -d tree functionality. The tree can be built by passing a set of points to the constructor. The additional functionality includes inserting new points using the function *insert*, resetting the tree using *reset*, re-balancing the tree with the function *reBalance*, and finding the  $k$  nearest neighbors of a given point using *findNearestNeighbors*. This functionality is enough for using the  $k$ -d tree in the emulation algorithm described in this paper. The current implementation uses the Euclidean distance, but it can be extended to include other metrics.

## B.2 Emulator

The emulator algorithm described in Section 2 has been implemented in the class *FastApproximator* in the include file `fast_approximator.hpp`. The constructor takes as input the dimensionality of the input and output spaces of the function  $f$  that is being approximated, the training set as two arrays of input and output values of  $f$ , and the number  $k$  of nearest neighbors to use in the approximation.<sup>10</sup> The approximation for a new input point can be performed with the *approximate* function, which takes as an input the new point, and the interpolation method (currently linear and quadratic interpolations are implemented), and returns as an output the approximated value of the function.<sup>11</sup> If desired, the function can also return the  $k$  nearest neighbors, distances to the nearest neighbors, and the indices of the nearest neighbors in the training set.

The function *approximate* is implemented in two steps, and each of the steps can be called separately as a function. The first step involves finding the  $k$  nearest neighbors, which is implemented in the function *findNearestNeighbors*. This function takes the new input point and finds the nearest neighbors. If desired, the nearest neighbors themselves, the distances to them, and their indices in the training set can also be returned. After this step, the function *getApproximation* can be called, which takes as an input the interpolation method and returns the approximated value. Note that this function can only be called after *findNearestNeighbors*. The separation of the approximation algorithm into two steps can be useful in many cases. For example, the error evaluation model, such as in Sect. 3, may only depend on the characteristics of the nearest neighbors. So the nearest neighbors can be found as a first step, the error can then be evaluated and only if the error is small enough for the approximation to be acceptable can the interpolation be performed. This approach can save computational time by not calculating the interpolation, which involves matrix operations and can be somewhat slow, unless it is necessary. Another useful application could be obtaining both linear and quadratic interpolations without repeating the nearest neighbor search. If none of this is necessary then the two step approximation can be replaced by a single call to the function *approximate*.

New points can be added to the training set with the function *addPoint*.<sup>12</sup> This function adds the new point to the  $k$ -d tree, without recalculating the covariance matrix, and checks if the  $k$ -d tree is very unbalanced after the insertion, in which case the tree is rebalanced by simply rebuilding it. The criterion for the tree being unbalanced is that the depth becomes 4 times larger than the depth of a balanced tree with the same number of elements (see Sect. 2).

The entire training set can be updated using the *reset* function, which takes as an input the new training set. The user can choose whether or not to update the covariance matrix of the input points. For example, if the new training set is a superset of the old one with not many additions, it may not be necessary to update the covariance matrix. Also, updating the covariance matrix will alter the distances to the already existing points, which may further impact the error evaluation model described below. So if one is simply adding a few new points to the training set without recalibrating the error model, then the covariance matrix should not be changed.

---

<sup>10</sup>The constructor is the equivalent of the function INITIALIZEEMUL in the schematic Alg. 1.

<sup>11</sup>The *approximate* function is equivalent to EMULATE in Alg. 1.

<sup>12</sup>Equivalent to ADDPOINTTOTRAINSET in Alg. 1.

### B.3 Error Model

The simple error model described in Sect. 3 has been implemented in the class *FastApproximatorError* in `fast_approximator_error.hpp`. The input to the constructor consists of a reference to a *FastApproximator* object, a set of test points (the cross-validation points), a function  $f_{\text{err}}$  for which the error is modelled, the *error modelling method*, the target accuracy, and the *decision method*.<sup>13</sup> We describe each of these parameters in detail below.

The function  $\eta : \mathbb{R}^m \rightarrow \mathbb{R}$  takes the output of  $f$  and maps it to a single number. This function is used for modelling the error (see Sect. 3). For the *Planck* results described in Sect. 6 the fast approximation algorithm outputs a set of  $C_l^{TT}$  values as well as a few of the *Planck* likelihoods, which can then be translated to the total *Planck* likelihood. In this case, the decision of whether or not to use the approximation should be based on the error of the total likelihood, rather than the errors of individual  $C_l$  values or individual likelihood components. The function  $\eta$  should then take the complex output of the approximation algorithm and translate it to the total likelihood.

As described in Sect. 3, the probability distribution  $p$  of the error  $e_\eta$  is modelled via a distribution that depends on the ratio of  $e_\eta$  to a single parameter  $\rho$  as in Eq. (3.3). The assumption is that the error increases linearly with  $\rho$ , where  $\rho$  can be calculated rapidly for any point in question and its nearest neighbors. Alg. 2 uses the average Euclidean distance to the nearest neighbors for  $\rho$ . The *error modelling method* for *FastApproximatorError* simply refers to the choice of  $\rho$  in general. The following options are currently implemented:

- *MIN\_DISTANCE* – The distance to the nearest neighbor.
- *AVG\_DISTANCE* – The average distance to the  $k$  nearest neighbors.
- *AVG\_INV\_DISTANCE* – The weighted average distance to the  $k$  nearest neighbors, with weights inversely proportional to the distances, which is useful when performing a weighted interpolation between the nearest neighbors.
- *SUM\_DISTANCE* – The geometrical sum of the distances to the nearest neighbors. This is a useful measure of both how far the nearest neighbors are on average and how isotropically they are distributed around the given point.
- *LIN\_QUAD\_DIFF* – The difference between linear and quadratic interpolations between the nearest neighbors.

The test set (also referred to as the cross-validation set  $X_{\text{CV}}$ ) is used to estimate the distribution of this ratio. For each point in the test set, the measure  $\rho$  is calculated, the approximate and exact values of  $\eta$  are calculated, and the ratio  $|e_\eta|/\rho$  is obtained. A histogram is then built using these ratios for all of the test points, which is further smoothed with a Gaussian kernel and normalized.

For any new point the parameter  $\rho$  is calculated, which is used to rescale the histogram above and obtain the distribution  $p$  for the error  $e_\eta$  at that point. The decision is made whether or not the approximation is acceptable, by translating the distribution to a single real number and comparing to the target threshold. The choice of the single number representing the distribution is made based on the *decision method* parameter passed to the constructor. The following choices are currently implemented:

---

<sup>13</sup>The constructor is equivalent to the `BUILDERRORMODEL` function in the schematic Alg. 2.



- *ONE\_SIGMA* – The 68.3% upper bound of the absolute value of the error.
- *TWO\_SIGMA* – The 95.5% upper bound of the absolute value of the error.
- *SQRT\_VAR* – The square root of the variance of the error.

The approximation for a new point is performed using the function *approximate*. The approximation itself is performed by passing the point to the underlying *FastApproximator*. The return value of this function indicates whether or not the approximation is acceptable. The function can also return the 68.3%, the 95.5% upper bounds of the error, as well as the variance and the mean, if any of those are requested.

The error model can be recalibrated with a new test set using the *reset* function. The *setPrecision* function can be used to change the target accuracy of the model (*i.e.*, the threshold for the error). The distribution for the ratio  $|e_\eta|/\rho$  can be obtained using the *getDistrib* function.

## B.4 Learn-As-You-Go

The *LearnAsYouGo* class in `learn_as_you_go.hpp` implements the “learn-as-you-go” functionality described in Sect. 4, using the *FastApproximator* and *FastApproximatorError* classes described above. The purpose of *LearnAsYouGo* is to decide for each input point whether or not it should approximate or do the calculation exactly. If the exact calculation is performed then the result is also added to the training set.

The input parameters to the constructor are the dimensionalities of the input and output spaces of the function  $f$  being calculated, a reference to  $f$ , the error model function  $\eta$ , the minimum number of points in the training set for which approximations are allowed, the error threshold  $\epsilon$ , as well as a file name in which the training set can be saved. The calculation is performed using the *evaluate* function,<sup>14</sup> which will return either an acceptable approximation for the value of  $f$  at the given point, or its exact value. If the exact value is calculated and returned then it is automatically added to the training set. The target threshold  $\epsilon$  can be changed at any time with the *setPrecision* function. The training set can be saved into a file using *writeIntoFile* and retrieved using *readFromFile*. The progress can be continuously logged into a file by setting the log file using the *logIntoFile* function. The log will include the total number of function calls, the number of calls for which the input point already exists in the training set, the number of calls with successful approximations, and the number of calls with unsuccessful approximations, *i.e.*, exact calculations.

An important characteristic of *LearnAsYouGo* is that it has a *parallel* implementation. If an application using *LearnAsYouGo* is run with many MPI processes, then each process will have its own “learn-as-you-go” approximation class. However, all of the processes will periodically share their training sets with each other, so effectively each process will use as a training set all of the exact function calculations by *all of the processes*. The user does not need to do anything special to take advantage of this functionality. All that is required is to run many parallel MPI processes.

## B.5 Accelerated *Planck* Likelihood

Finally, we use the *LearnAsYouGo* class to implement an accelerated “learn-as-you-go” *Planck* likelihood in the *PlanckLikeFast* class in `planck_like_fast.hpp`. The interface is

---

<sup>14</sup>Equivalent to `CALCULATELAYG` in Alg. 3.



similar to the *Planck* likelihood interface in COSMO++ [25]. The implementation uses the *LearnAsYouGo* class with the  $C_l^{TT}$  values and the slower *Planck* likelihoods as the function to be approximated (see Sect. 6), and the total *Planck* likelihood function to model the error. All of the functionality of *LearnAsYouGo* described above, including the parallel implementation, is present in *PlanckLikeFast*. Examples of using *PlanckLikeFast* in parameter space samplers are implemented in the `test_mcmc_planck_fast.cpp` and `test_multinest_planck_fast.cpp` files for Metropolis-Hastings and MULTINEST samplers, respectively. Those implementations are used to obtain the results in Sect. 6.

The constructor takes as an input a pointer to a *CosmologicalParams* object, which is used for setting the cosmological parameters, boolean (logical) values defining which of the likelihoods need to be included and whether or not tensor modes are included, as well as the number of points per decade in  $k$  space for which the primordial power spectrum is evaluated. Following these parameters, two parameters for the “learn-as-you-go” algorithm must be specified. These parameters are the target accuracy (the threshold for the error used to decide whether or not the approximation is acceptable), and the minimum size of the training set before approximations are performed.

The calculation is done through the *calculate* function. This function enables the use of the *PlanckLikeFast* class in parameter space samplers. Since parameter space sampling is the main application for the “learn-as-you-go” likelihood, the detailed interface for calculating each component of the likelihood separately is not implemented here. The regular *Planck* likelihood interface *PlanckLikelihood* can be used for that purpose as in Ref. [25]. With this interface, it is very easy to replace the usual *Planck* likelihood code by the accelerated one in any parameter space sampling code. Simply replacing the *PlanckLikelihood* object with a *PlanckLikeFast* object initialized with all of the necessary parameters will suffice to use the accelerated likelihood code in any parameter space sampler present in COSMO++.

Since the function being approximated in this case is a likelihood function, we can apply the techniques developed in Sect. 5 to estimate the errors of the posterior distributions for the parameters of interest. Our implementation of *PlanckLikeFast* allows easy calculation of the posteriors *with errors* as follows. Before starting the parameter space sampling an error log file must be set using the *logError* function of *PlanckLikeFast*. Once this is set every call of *calculate* will output the call information into the error log file, which includes the parameters for the function call, the resulting likelihood, and estimated characteristics of the error, which can then be used to estimate the errors of the posteriors. Whenever the *calculate* function does the exact calculation the errors will be zero. There is no modification to the chain files produced by the parameter space samplers. After the sampling is finished, the *MarkovChain* class can be used to analyze the chains and obtain the posteriors. The new version of *MarkovChain* allows for calculation of errors of the posteriors by simply passing the error log file produced by *PlanckLikeFast* as an extra argument to the constructor. In summary, all that is needed for producing posteriors with errors is generating an error log file in *PlanckLikeFast* and then passing it to *MarkovChain* for analyzing.

## References

- [1] **WMAP** Collaboration, G. Hinshaw *et. al.*, *Nine-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results, Astrophys.J.Suppl.* **208** (2013) 19, [[arXiv:1212.5226](#)].

- [2] **Planck** Collaboration, P. Ade *et. al.*, *Planck 2013 results. XV. CMB power spectra and likelihood*, *Astron.Astrophys.* **571** (2014) A15, [[arXiv:1303.5075](#)].
- [3] A. Lewis, A. Challinor, and A. Lasenby, *Efficient computation of CMB anisotropies in closed FRW models*, *Astrophys.J.* **538** (2000) 473–476, [[astro-ph/9911177](#)].
- [4] J. Lesgourgues, *The Cosmic Linear Anisotropy Solving System (CLASS) I: Overview*, [[arXiv:1104.2932](#)].
- [5] D. Blas, J. Lesgourgues, and T. Tram, *The Cosmic Linear Anisotropy Solving System (CLASS) II: Approximation schemes*, *JCAP* **1107** (2011) 034, [[arXiv:1104.2933](#)].
- [6] A. O’Hagan, *Bayesian analysis of computer code outputs: A tutorial*, *Reliability Engineering & System Safety* **91** (2006), no. 1011 1290 – 1300.
- [7] M. Kaplinghat, L. Knox, and C. Skordis, *Rapid calculation of theoretical cmb angular power spectra*, *Astrophys.J.* **578** (2002) 665, [[astro-ph/0203413](#)].
- [8] R. Jimenez, L. Verde, H. Peiris, and A. Kosowsky, *Fast cosmological parameter estimation from microwave background temperature and polarization power spectra*, *Phys.Rev.* **D70** (2004) 023005, [[astro-ph/0404237](#)].
- [9] C. Ringeval, *Fast Bayesian inference for slow-roll inflation*, *Mon.Not.Roy.Astron.Soc.* **439** (2014) 3253, [[arXiv:1312.2347](#)].
- [10] A. Petri, J. Liu, Z. Haiman, M. May, L. Hui, *et. al.*, *Emulating the CFHTLenS Weak Lensing data: Cosmological Constraints from moments and Minkowski functionals*, [[arXiv:1503.0621](#)].
- [11] W. Fendt, J. Chluba, J. Rubino-Martin, and B. Wandelt, *Rico: An Accurate Cosmological Recombination Code*, *Astrophys.J.Suppl.* **181** (2009) 627–638, [[arXiv:0807.2577](#)].
- [12] W. A. Fendt and B. D. Wandelt, *Pico: Parameters for the Impatient Cosmologist*, *Astrophys.J.* **654** (2006) 2–11, [[astro-ph/0606709](#)].
- [13] W. A. Fendt and B. D. Wandelt, *Computing High Accuracy Power Spectra with Pico*, *Astrophys.J.* (2007) [[arXiv:0712.0194](#)].
- [14] A. Bouland, R. Easther, and K. Rosenfeld, *Caching and Interpolated Likelihoods: Accelerating Cosmological Monte Carlo Markov Chains*, *JCAP* **1105** (2011) 016, [[arXiv:1012.5299](#)].
- [15] K. Heitmann, M. White, C. Wagner, S. Habib, and D. Higdon, *The Coyote Universe I: Precision Determination of the Nonlinear Matter Power Spectrum*, *Astrophys.J.* **715** (2010) 104–121, [[arXiv:0812.1052](#)].
- [16] K. Heitmann, D. Higdon, M. White, S. Habib, B. J. Williams, *et. al.*, *The Coyote Universe II: Cosmological Models and Precision Emulation of the Nonlinear Matter Power Spectrum*, *Astrophys.J.* **705** (2009) 156–174, [[arXiv:0902.0429](#)].
- [17] E. Lawrence, K. Heitmann, M. White, D. Higdon, C. Wagner, *et. al.*, *The Coyote Universe III: Simulation Suite and Precision Emulator for the Nonlinear Matter Power Spectrum*, *Astrophys.J.* **713** (2010) 1322–1331, [[arXiv:0912.4490](#)].
- [18] M. D. Schneider, O. Holm, and L. Knox, *Intelligent Design: On the Emulation of Cosmological Simulations*, *Astrophys.J.* **728** (2011) 137, [[arXiv:1002.1752](#)].
- [19] J. Kwan, S. Bhattacharya, K. Heitmann, and S. Habib, *Cosmic Emulation: The Concentration-Mass Relation for  $w$ CDM Universes*, *Astrophys.J.* **768** (2013) 123, [[arXiv:1210.1576](#)].
- [20] J. Kwan, K. Heitmann, S. Habib, N. Padmanabhan, H. Finkel, *et. al.*, *Cosmic Emulation: Fast Predictions for the Galaxy Power Spectrum*, [[arXiv:1311.6444](#)].
- [21] K. Heitmann, E. Lawrence, J. Kwan, S. Habib, and D. Higdon, *The Coyote Universe Extended: Precision Emulation of the Matter Power Spectrum*, *Astrophys.J.* **780** (2014) 111,

- [[arXiv:1304.7849](#)].
- [22] T. Auld, M. Bridges, M. Hobson, and S. Gull, *Fast cosmological parameter estimation using neural networks*, *Mon.Not.Roy.Astron.Soc.Lett.* **376** (2007) L11–L15, [[astro-ph/0608174](#)].
  - [23] P. Graff, F. Feroz, M. P. Hobson, and A. Lasenby, *BAMBI: blind accelerated multimodal Bayesian inference*, *Mon.Not.Roy.Astron.Soc.* **421** (2012) 169–180, [[arXiv:1110.2997](#)].
  - [24] P. Graff, F. Feroz, M. P. Hobson, and A. N. Lasenby, *SKYNET: an efficient and robust neural network training tool for machine learning in astronomy*, *Mon.Not.Roy.Astron.Soc.* **441** (2014), no. 2 1741–1759, [[arXiv:1309.0790](#)].
  - [25] G. Aslanyan, *Cosmo++: An Object-Oriented C++ Library for Cosmology*, *Comput.Phys.Commun.* **185** (2014) 3215–3227, [[arXiv:1312.4961](#)].
  - [26] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo In Practice*. Chapman and Hall, 1996.
  - [27] F. Feroz and M. Hobson, *Multimodal nested sampling: an efficient and robust alternative to MCMC methods for astronomical data analysis*, *Mon.Not.Roy.Astron.Soc.* **384** (2008) 449, [[arXiv:0704.3704](#)].
  - [28] F. Feroz, M. Hobson, and M. Bridges, *MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics*, *Mon.Not.Roy.Astron.Soc.* **398** (2009) 1601–1614, [[arXiv:0809.3437](#)].
  - [29] F. Feroz, M. Hobson, E. Cameron, and A. Pettitt, *Importance Nested Sampling and the MultiNest Algorithm*, [arXiv:1306.2144](#).
  - [30] J. L. Bentley, *Multidimensional binary search trees used for associative searching*, *Commun. ACM* **18** (Sept., 1975) 509–517.
  - [31] P. Billingsley, *Probability and Measure*. John Wiley & sons, third ed., 1995.
  - [32] A. Lewis and S. Bridle, *Cosmological parameters from CMB and other data: A Monte Carlo approach*, *Phys.Rev.* **D66** (2002) 103511, [[astro-ph/0205436](#)].
  - [33] A. Gelman and D. B. Rubin, *Inference from iterative simulation using multiple sequences*, *Statistical science* (1992) 457–472.